Identifying and Characterizing a *Consumer Medical Vocabulary*

A Dissertation Research Project
by

Tony Tse

Under the Direction
of

Dagobert Soergel

Ultimate Objective

Enhancing healthcare consumer accessibility to and comprehension of medical information.

“Even though nothing can substitute for the expertise of your own doctor, no prescription is more valuable than knowledge.”

- Dr. C. Everett Koop

Outline

- Problem
- Conceptual Framework
- Background
- Methodology
- Results
- Conclusions
- Implications

Problem

- Medical domain access by non-healthcare professionals
  - Information mediators (“mediators”)
  - Healthcare consumers (“consumers”)

- Barriers
  - Terminology Gap: *What is it called? What does it mean?*

    “breathing difficulty” = “dyspnea”

  - Conceptualization: *How does it work?*

    mechanisms of “autoimmune response”

General Objectives

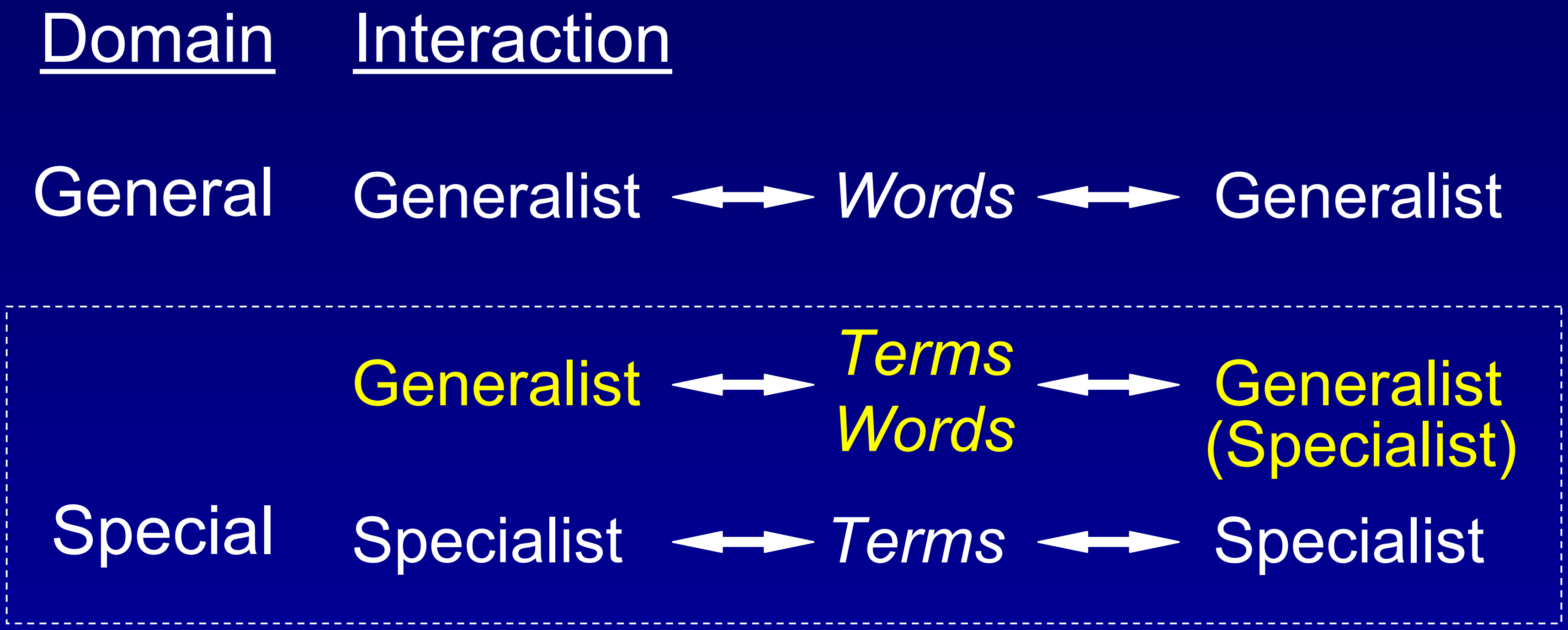
- Characterize terms used by non-professionals to describe medical concepts
  - Compare with professional medical terms
  - Compare within certain context
    (e.g., disease duration)
- Develop and evaluate procedures for corpus-based extraction and analysis of terms used by non-professionals

Research Questions

- What is a Consumer Medical Vocabulary (CMV)? Is terminological theory a viable model?
- Does a Mediator Medical Vocabulary (MMV) bridge medical vocabularies used by professionals (PMV) and consumers (CMV)?
- Do vocabularies differ in "expressive variability?" Do the most frequently used forms for a concept ("consensus forms") differ by vocabulary?

Conceptual Framework - Terminology

- Terminology as a conceptual interface
  - Words: General discourse
  - Terms: Specialized discourse



Domain	Interaction				
General	Generalist	↔	*Words*	↔	Generalist
	Generalist	↔	*Terms* *Words*	↔	Generalist (Specialist)
Special	Specialist	↔	*Terms*	↔	Specialist

Conceptual Framework - Communication

- Terminology and access
  - Form: Surface-level structure
  - Concept: Deep structure

		Concept	
		Understood	Not Understood
Form	Known	**Communication**	**Mis-communication**
	Unknown	**Mis-communication**	**No Communication**

Background - Examples of Terms

- Term = <Form, Concept>

  <antacid, C0003138 (Antacids)>

- Concept: Unified Medical Language System® (UMLS)

  C0027051 (Myocardial Infarction)

  Unique ID Preferred Term

- Example of synonyms and homonyms

  <hypersensitivity, C0020517 (Hypersensitivity)>
  <allergy, C0020517 (Hypersensitivity)>

  <depression, C0011570 (Mental Depression)>
  <depression, C0497301 (Feeling Depressed)>

Background - Term Relationships

Relationship (Term1 ↔ Term 2)	Form	Concept
Identical	Same	Same
Synonym	Different	Same
Homonym	Same	Different
Unique	Different	Different

Background - Term Standardization

- Standardization facilitates comparison
- Form
  - String normalization

    heart attack, Heart Attacks → heart attack
    colonic cancer, cancer of the colon → colon cancer
  - Non-regular forms

    abbreviations, acronyms, clippings → expanded
    coordinate constructs (“head and neck injury”) →
    two forms (“head injury” & “neck injury”)
- Concept: UMLS concept unique ID + preferred term

Background - Concept Semantic Types

- UMLS Semantic Network
  - Concepts classified by semantic types
  - Semantic types linked by semantic relations
- Semantic Types (134)

  C0027051 (myocardial infarction) → Disease or Syndrome
  C0018681 (headache) → Sign or Symptom
  C0392806 (hip replacement) → Therapeutic or Preventative Procedure

- Semantic Type Groups (15)

  Disorders (PATH): Disease or Syndrome + Sign or Symptom + …

  Procedures (PROC): Therapeutic or Preventative Procedure + …

  Anatomy (ANAT)

  Chemicals & Drugs (CHEM)

Background - Mapping to UMLS

- Match term with UMLS concept that “best” represents its meaning
  - Close: synonyms, quasi-synonyms

    “lump” → C0024873 (Mass, NOS)
  - Approximate: hyponyms, hypernyms

    “large lump” **N**→ C0024873 (Mass, NOS)
- Homonyms: Context dependent

  “diet” → C0600072 (Feeding and dietary regimens)
  “diet” → C0012155 (Diet)

Methodology - Approach

- Corpus-based terminography
  - Documents authored by laypersons
  - “Utterances” reviewed in context
  - Forms mapped in context
- Manual term extraction from lay perspective
  - Labor intensive, but increases “authenticity”
  - Identification of “free phrases,” idioms, slang, other “regular forms” with medical connotations
  - Identification of “non-regular forms” such as acronyms, abbreviations, clippings, typos
- Frequently-occurring usage patterns

Methodology - Procedure Overview

- Corpus Generation
  - Document source selection
  - Document selection
- Vocabulary Generation
  - Term extraction
  - Form processing
  - Mapping terms to UMLS Concepts
- Analysis of Vocabulary Characteristics
  - Form-based characteristics
  - Concept-based characteristics
  - Form-concept relationships
  - Term-based characteristics

Methodology - Corpus Generation

- Sources of Terms
  - Consumer: Web-based discussion forum posts
  - Mediator: Magazine & newspaper articles, ads
  - Professional: MeSH and SNOMED (controlled terms)
- Privacy/Copyright
  - Consumer: IRB-approved exemption
  - Mediator: Fair use rules
- Selection Criteria: Max breadth of scope
- Corpus Size
  - Consumer: 1,900 postings; 25,000 forms
  - Mediator: 500 articles; 21,300 forms

Methodology - Vocabulary Generation 1

- Term Extraction
  - 14 “consumer surrogates” were trained
  - Identified terms, but extracted **forms**
  - 2 extractors reviewed each document
    - 55% complete form overlap
    - 22% partial form overlap
- Form Processing
  - Researcher reviewed forms (~6% modified)
  - Form normalization: Expansion, spelling….
- Form Mapping: MetaMap & manual process

Methodology - Vocabulary Generation 2

"I had a heart atack two years ago, but the Heart Doc says I'm O.K. based on EKG."

Extracted Forms	Preprocessed Forms
heart atack	heart attack
Heart Doc	heart doctor
O.K.	ok
EKG	electrocardiogram

Methodology - Vocabulary Generation 3

- Mapping to UMLS

Mapped Forms	Mapped-to UMLS Concepts
heart attack	→ C0027051 (Myocardial Infarction)
heart doctor	→ C0175906 (Cardiologist)
ok	N → C0018684 (Health)
electrocardiogram	→ C0013798 (Electrocardiogram)

	CMV	MMV
• Portion of forms mapped:	99%	92%

Methodology - Vocabulary Size

	CMV	MMV
Form Tokens:	55,000	45,800
Form Types:	25,000	21,300
Concept Tokens:	54,500	42,100
Concept Types:	5,300	5,400

Results - Findings

- Differences at the form level between health professionals and non-professionals

- Few summary differences among vocabularies (CMV, MMV, and PMV)

- Non-professional terms are highly context-sensitive

Results - Form Level

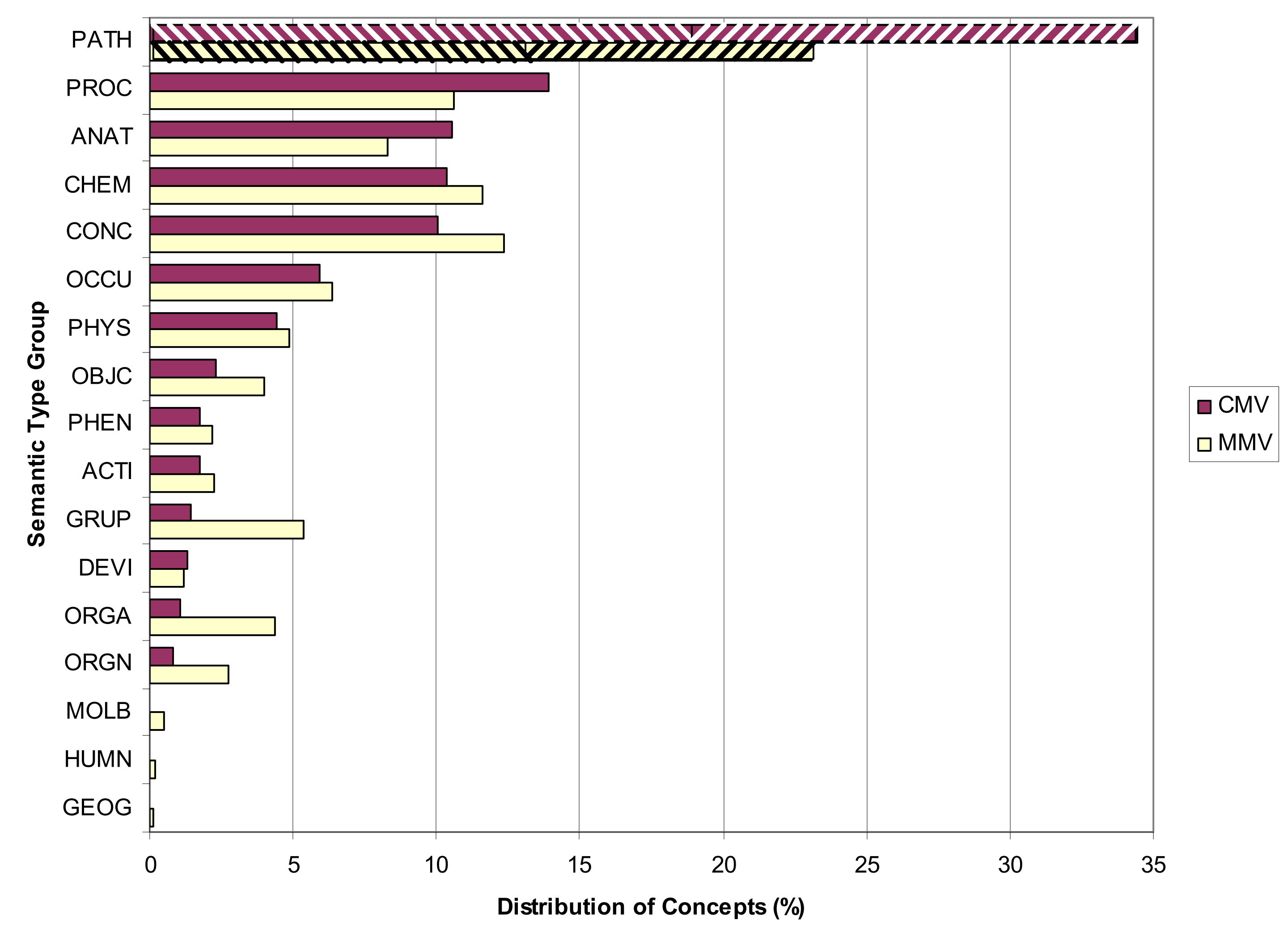
- Mean length: CMV < MMV < PMV

Normalized chars	16.8	18.2	23.5
Words	2.2	2.2	2.4

- Areas of 30 most frequent forms in each vocab (16% CMV & 16% MMV by Token)
  - Top 3 areas by token
    - CMV: General discourse, symptoms, anatomy
    - MMV: General discourse, epidemiology-populations, research methodology
  - Unique to MMV

    epidemiology-populations, research methodology

Results - Concept Level



Disorders (PATH) Clusters: \\\ - “disease”; /// - “symptom”

Results – Expressive Variability

- Operationalized as forms per concept
- Limited to closely mapped-to UMLS concepts
- 20 concepts in each vocab with most variability
  - Overall: “subjective” > “objective” concepts
  - Number of “subjective” concepts: CMV > MMV

C0683369
(Clouded Consciousness)
space out
fog
spaciness
mind was in a fog
zombie

C0013231
(Drug, Non-Prescription)
over the counter {medication, drug}
otc { }
nonprescription { }

Results – Consensus Form

- Few forms account for over 50% of concept occurrences
- Overlap of consensus forms with PMV forms: MMV > CMV

CMV	MMV
diagnosis (90%)	side effect (85%)*
side effect (88%)*	control (76%)
health (54%)	infect (75%)
treatment (53%)	high risk (54%)

* Sense: Injury or Poisoning

Results - Vocabulary Overlap

- Pair-wise vocabulary comparison (one-sided)
  - Closely mapped-to concepts only
- Non-professional → Professional
  - Conceptual overlap: 80%
  - Form commonality
    - Complete: 55%
    - Partial: 18%
    - None: 27%
- Non-professional → Non-professional
  - Conceptual overlap: 48%

Results - Research Questions

- Existence of CMV depends on definition
  - “Terms” used by laypersons in medical domain
  - “Terms” used only by laypersons, distinct from both general and special domains
- Terminology = viable model/process
- MMV “bridging” function not observed
- Consensus forms = “common level of discourse” within groups

Limitations

- Validity (e.g., mapping, comprehension)
- Genre “mismatch” (CMV vs. MMV)
- Breadth of topics (e.g., duration)
- Reliability (e.g., coding “drift”)
- PMV: controlled, not extracted, terms
- Forms only: not terms from extractors
- No pragmatics: “anthrax is a virus”

Implications

- Preliminary data about characteristics of medical terms used by non-professionals
  - Automated extraction: String probes
  - Interface design: Contextualization
  - Theory: Generalists in specialist domains
- Non-professional forms and UMLS concepts
  - Readability research
  - Thesaurus/entry vocabulary for consumers
- Procedure for exploring terms bordering general and special domains

Future Research

- Exploratory research → insights/methods need to be validated (e.g., expert review)
- Scale-up: algorithms/heuristics (automation)
- Field studies to understand conceptual systems/mental models of consumers (comprehension)
- Analysis at the pragmatic level

Acknowledgements

Advisory Committee

Professor Dagobert Soergel, Chair/Advisor
Professor Eileen G. Abels
Dr. Keith W. Cogdill
Professor Linda K. Coleman
Professor Gary Marchionini